CS 466
Introduction to Bioinformatics

Instructor: Jian Peng
Teaching Assistant: Wesley Qian
Random Variable

Quite commonly, we would like to deal with numbers that are random. We can do so by linking numbers to the outcome of an experiment. We define a random variable:

**Definition: 4.1 Discrete random variable**

Given a sample space $\Omega$, a set of events $\mathcal{F}$, a probability function $P$, and a countable set of real numbers $D$, a discrete random variable is a function with domain $\Omega$ and range $D$. 

\[
X = \begin{cases} 
0 \\
1
\end{cases}
\]
**Definition: 4.8 Expected value**

Given a discrete random variable $X$ which takes values in the set $\mathcal{D}$ and which has probability distribution $P$, we define the expected value

$$\mathbb{E}[X] = \sum_{x \in \mathcal{D}} xP(X = x).$$

This is sometimes written $\mathbb{E}_P[X]$, to clarify which distribution one has in mind.

**Example: 4.5 Betting on coins**

We agree to play the following game. I flip a fair coin (i.e. $P(H) = P(T) = 1/2$). If the coin comes up heads, you pay me 1; if the coin comes up tails, I pay you 1. The expected value of my income is 0, even though the random variable never takes that value.
Definition: 4.9 \textit{Expectation}

Assume we have a function $f$ that maps a discrete random variable $X$ into a set of numbers $D_f$. Then $f(X)$ is a discrete random variable, too, which we write $F$. The expected value of this random variable is written

$$E[f] = \sum_{u \in D_f} uP(F = u) = \sum_{x \in D} f(x)P(X = x)$$

which is sometimes referred to as “the expectation of $f$”. The process of computing an expected value is sometimes referred to as “taking expectations”.

Definition: 4.10 \textit{Expected value of a continuous random variable}

Given a continuous random variable $X$ which takes values in the set $D$ and which has probability distribution $P$, we define the expected value

$$E[X] = \int_{x \in D} xp(x)dx.$$ 

This is sometimes written $E_p[X]$, to clarify which distribution one has in mind.
Mean, Variance and Covariance

Definition: 4.12  Mean or expected value
The mean or expected value of a random variable $X$ is

$$\mathbb{E}[X]$$

Definition: 4.13  Variance
The variance of a random variable $X$ is

$$\text{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Definition: 4.14  Covariance
The covariance of two random variables $X$ and $Y$ is

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$
Examples

Worked example 4.9  Mean of a coin flip

We flip a biased coin, with $P(H) = p$. The random variable $X$ has value 1 if the coin comes up heads, 0 otherwise. What is the mean of $X$? (i.e. $\mathbb{E}[X]$).

Solution: $\mathbb{E}[X] = \sum_{x \in \{0, 1\}} xP(X = x) = 1p + 0(1 - p) = p$

Worked example 4.10  Variance of a coin flip

We flip a biased coin, with $P(H) = p$. The random variable $X$ has value 1 if the coin comes up heads, 0 otherwise. What is the variance of $X$? (i.e. $\text{var}[X]$).

Solution: $\text{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = (1p - 0(1 - p)) - p^2 = p(1 - p)$

Worked example 4.11  Variance

Can a random variable have $\mathbb{E}[X] > \sqrt{\mathbb{E}[X^2]}$?

Solution: No, because that would mean that $\mathbb{E}[(X - \mathbb{E}[X])^2] < 0$. But this is the expected value of a non-negative quantity; it must be non-negative.
Properties of variance and covariance

Useful Facts: 4.3  \textit{Properties of variance}

1. For any constant $k$, $\text{var}[k] = 0$
2. $\text{var}[X] \geq 0$
3. $\text{var}[kX] = k^2 \text{var}[X]$
4. if $X$ and $Y$ are independent, then $\text{var}[X + Y] = \text{var}[X] + \text{var}[Y]$
5. $\text{var}[X] = \text{cov}(X, X)$.

1, 2, and 5 are obvious. You will prove 3 and 4 in the exercises.

Useful Facts: 4.6  \textit{Independent random variables have zero covariance}

1. if $X$ and $Y$ are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
2. if $X$ and $Y$ are independent, then $\text{cov}(X, Y) = 0$.

If 1 is true, then 2 is obviously true (apply the expression of useful facts\textit{4.5}). I prove 5 below.
Properties

Useful Facts: 4.4  A useful expression for variance

\[
\text{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]
= \mathbb{E}\left[(X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2)\right]
= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2
= \mathbb{E}[X^2] - (\mathbb{E}[X])^2
\]


Useful Facts: 4.5  A useful expression for covariance

\[
\text{cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]
= \mathbb{E}[(XY - Y\mathbb{E}[X] - X\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y])]
= \mathbb{E}[XY] - 2\mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]
= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].
\]
Proposition: If $X$ and $Y$ are independent random variables, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Proof: Recall that $\mathbb{E}[X] = \sum_{x \in D} xP(X = x)$, so that

$$
\mathbb{E}[XY] = \sum_{(x,y) \in D_x \times D_y} xyP(X = x, Y = y)
= \sum_{x \in D_x} \sum_{y \in D_y} (xyP(X = x, Y = y))
= \sum_{x \in D_x} \sum_{y \in D_y} (xP(X = x))P(Y = y))
$$

because $X$ and $Y$ are independent

$$
= \left( \sum_{x \in D_x} xP(X = x) \right) \left( \sum_{y \in D_y} yP(Y = y) \right)
= (\mathbb{E}[X])(\mathbb{E}[Y]).
$$

This is certainly not true when $X$ and $Y$ are not independent (try $Y = -X$).
Statistics
One simple and effective summary of a set of data is its **mean**. This is sometimes known as the **average** of the data.

**Definition: 1.1  Mean**

Assume we have a dataset \( \{x\} \) of \( N \) data items, \( x_1, \ldots, x_N \). Their mean is

\[
\text{mean} ( \{x\} ) = \frac{1}{N} \sum_{i=1}^{i=N} x_i.
\]
Definition: 1.2 Standard deviation

Assume we have a dataset \( \{x\} \) of \( N \) data items, \( x_1, \ldots, x_N \). The standard deviation of this dataset is:

\[
\text{std} (\{x_i\}) = \sqrt{\frac{1}{N} \sum_{i=1}^{i=N} (x_i - \text{mean}(\{x\}))^2} = \sqrt{\text{mean}(\{(x_i - \text{mean}(\{x\}))^2\})}.
\]

Definition: 1.3 Variance

Assume we have a dataset \( \{x\} \) of \( N \) data items, \( x_1, \ldots, x_N \). where \( N > 1 \). Their variance is:

\[
\text{var} (\{x\}) = \frac{1}{N} \left( \sum_{i=1}^{i=N} (x_i - \text{mean}(\{x\}))^2 \right) = \text{mean}(\{(x_i - \text{mean}(\{x\}))^2\})
\]
Normalization

\[ \hat{x}_i = \frac{(x_i - \text{mean} \{x\})}{\text{std} \{x\}}. \]
FIGURE 2.16: The three kinds of scatter plot are less clean for real data than for our idealized examples. Here I used the body temperature vs heart rate data for the zero correlation; the height-weight data for positive correlation; and the lynx data for negative correlation. The pictures aren’t idealized — real data tends to be messy — but you can still see the basic structures.
Definition: 2.1 Correlation coefficient

Assume we have $N$ data items which are 2-vectors $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. These could be obtained, for example, by extracting components from larger vectors. We compute the correlation coefficient by first normalizing the $x$ and $y$ coordinates to obtain $\hat{x}_i = \frac{(x_i - \text{mean}\{x\})}{\text{std}(x)}$, $\hat{y}_i = \frac{(y_i - \text{mean}\{y\})}{\text{std}(y)}$. The correlation coefficient is the mean value of $\hat{x}\hat{y}$, and can be computed as:

$$\text{corr} \left( \{(x, y)\} \right) = \frac{\sum_i \hat{x}_i \hat{y}_i}{N}$$

Also called **Pearson Correlation Coefficient**
Correlation coefficient vs Relationship
Correlation and Causality

Ice Cream vs Drowning

Drowning Deaths

Consumption of Ice-Cream

monthly data

Ice Cream vs Drowning
Ice Cream vs Drowning

Graph showing the comparison between ice cream consumption and drownings over the months of the year.

- **Ice cream consumption**
  - Highest consumption in July and August
  - Moderate consumption in June and September
  - Lowest consumption in January and December

- **Drownings**
  - Highest numbers in July and August
  - Moderate numbers in June and September
  - Lowest numbers in January and December
Chocolate vs Nobel Prizes

$r = 0.791$
$P < 0.0001$

credit: NEJM, 2012
Gene expression analysis

Correlation of genes across experimental conditions → coregulation of genes
Correlation analysis

\[
r = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2 \sum(Y - \overline{Y})^2}}
\]

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>...</th>
<th>Sample n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>(X_{11})</td>
<td>(X_{12})</td>
<td>...</td>
<td>(X_{1n})</td>
</tr>
<tr>
<td>Gene 2</td>
<td>(X_{21})</td>
<td>(X_{22})</td>
<td>...</td>
<td>(X_{2n})</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Gene m</td>
<td>(X_{m1})</td>
<td>(X_{m2})</td>
<td>...</td>
<td>(X_{mn})</td>
</tr>
</tbody>
</table>

Gene 1: \(r = -0.8\)  
Gene 2: \(r = -0.2\)  
Gene 3: \(r = 0.85\)  
Gene m: \(r = -0.15\)